Syllable-Length Acoustic Units in Large-Vocabulary Continuous Speech Recognition

Annika Hämäläinen, Lou Boves, Johan de Veth

Centre for Language and Speech Technology (CLST)
Radboud University Nijmegen, Nijmegen, The Netherlands
{A.Hamalainen, L.Boves, J.deVeth}@let.ru.nl

Abstract

Recent research on the TIMIT corpus suggests that longer-length acoustic units are better suited for modelling coarticulation and long-term temporal dependencies in speech than conventional context-dependent phone models. However, the impressive results achieved on TIMIT [1] are yet to be reproduced on other corpora, such as read speech from the Spoken Dutch Corpus. Differences between TIMIT and the Spoken Dutch Corpus data are analysed in an attempt to better understand in which conditions the use of longer-length units can be expected to result in considerable improvements in recognition accuracy. We conclude that at least part of the improvements found with TIMIT can be explained by details of the experimental procedure, and that longer-length left-to-right HMMs that borrow their topology from a sequence of triphones are only able to capture part of the pronunciation variation present in speech.

1. Introduction

Longer-length acoustic units, based e.g. on words or syllables, have been suggested as an attractive alternative for conventionally used context-dependent phone models, such as triphones, in large-vocabulary continuous speech recognition (LVCSR) [1-4]. This is because they are expected to better capture long-term spectral and temporal dependencies related to pronunciation variation and coarticulation than short, phoneme-length units.

Promising results with longer-length acoustic units have recently been published [1-4]. [1] proposed a hierarchical method, which employs a mixture of word-, syllable- and phoneme-based units. As the gain reported in recognition accuracy was particularly impressive, the method was adopted for our research, the results of which were reported in [2]. [1] uses TIMIT, a database which contains carefully read and annotated American English, and has specifically been designed for recognition experiments. To validate our implementation of the method, the recognition experiments on TIMIT were repeated. In addition, similar experiments were carried out for another language and speech style, viz. Dutch speech read in a more lively style – and equipped with a coarser annotation. As in the case of [3, 4], the improvements gained were more modest than those achieved on TIMIT. Somewhat surprisingly, none of the above-mentioned papers make an attempt to explain why the excellent results obtained on TIMIT cannot be replicated on other corpora. The aim of this paper is to fill that gap: we seek to shed light on why results achieved on TIMIT overestimate the gain that can be obtained with longer-length acoustic units on other corpora.

This paper is further organised as follows. The speech material used in our recognition experiments is described in Section 2. The experimental set-up is detailed in Section 3. The results of a first set of experiments are presented in Section 4, followed by analyses and the results of a second set of experiments in Section 5. Our findings are discussed in Section 6 and, finally, our conclusions are formulated in Section 7.

2. Speech material

TIMIT [5] is manually labelled and includes time-aligned, manually verified phone and word segmentations. For this study, the original set of labels was reduced to a set of 45 phone labels. The Dutch speech material was extracted from the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) [6]. The CGN data used in this study were read speech from a library for the blind; they comprise manually verified (broad) phonetic and word labels, as well as manually verified word-level segmentations. A set of 37 phone labels was used for CGN. Apart from the languages and the labelling and segmentation protocols, the main difference between the two data sets is the more lively nature of the Dutch speech – related to its purpose to be used for entertainment.

The data for each language were divided into three sets: a set for training the acoustic models, a test set for evaluating the acoustic models, and a development set for optimising the minimum number of training examples (see Subsection 3.3), language model scaling factor, and word insertion penalty. Details of the data are presented in Tables 1 and 2.

Table 1: TIMIT data sets.

	Train	Test	Devel.	Total
Word tokens	30,132	9,455	1,570	41,157
Speakers	462	144	24	630
hh:mm:ss	3:08:42	0:59:13	0:09:43	4:17:38

Table 2: CGN data sets.

	Train	Test	Devel.	Total
Word tokens	45,172	7,917	7,507	60,596
Speakers	125	125	125	125
hh:mm:ss	4:51:27	0:51:34	0:48:13	6:31:14

3. Experimental set-up

3.1. Feature extraction

Feature extraction was carried out at a frame rate of 10 ms, with a pre-emphasis of 0.97. 12 Mel Frequency Cepstral

Coefficients and log-energy with first and second order time derivatives were calculated, for a total of 39 features. Channel normalisation was applied using cepstral mean normalisation over individual sentences for TIMIT and complete recordings (with a mean duration of 3.5 minutes) for CGN. Feature extraction was performed using HTK [7].

3.2. Lexica and language models

The research in this paper addresses issues in acoustic modelling. In order to study possible improvements due to changes in acoustic modelling only – without the risk of language modelling issues masking the effects – out-of-vocabulary words were not allowed in the tasks. In effect, the recognition lexicon and word-level bigram network for each language were built using all orthographic words in the training and test sets. The vocabulary consisted of about 6,000 words for TIMIT and 10,500 words for CGN. The test set perplexity, computed on a per-sentence basis using HTK [7], was 16 for TIMIT and 46 for CGN. These numbers reflect the inherent differences between the corpora and the resulting recognition tasks.

3.3. Acoustic modelling

In [2] we showed that, for both TIMIT and CGN, the best trade-off between recognition performance and the number of model parameters is obtained with a mix of syllable units (including monosyllabic words) and triphones. In preparation for building a mixed-unit recogniser that employs syllable units and triphones, two recognisers were built: a triphone and a syllable-unit recogniser.

A standard procedure with decision tree state tying was used to train the triphone recogniser [7]. As opposed to [1], which used the manual labels in combination with a flat start Baum-Welch re-estimation strategy, the TIMIT triphones were bootstrapped using the manually verified phonetic segmentations of the sentences. Since CGN has manually verified word segmentations, in addition to manually verified broad phonetic labels, initial 32-Gaussian monophones were trained using linear segmentation within the word segments. The monophones were used to perform a forced alignment of the training data; the CGN triphones were then bootstrapped using the resulting phone segmentations.

The context-free syllable units of the syllable-unit recogniser were initialised with the 8-Gaussian triphone models corresponding to the underlying (canonical) phonemes of the syllables. To capture the spectral and temporal dependencies, the syllable units that appeared frequently enough in the training data were trained further using Baum-Welch re-estimation. The optimal set of further trained units was determined by experimenting with different values for the minimum number of tokens required for the further training of a unit. Performance on the development test set was used as the criterion. Only robustly trained units from the syllable-unit recogniser were used in the mixed-unit recogniser; when syllables did not occur frequently enough in the training data, triphones were backed off to. The mix of units underwent four more passes of Baum-Welch re-estimation.

Recognition experiments on TIMIT and CGN were carried out using all three types of recognisers. The baseline performance was determined by the performance of the triphone recogniser.

4. Results

4.1. TIMIT

The results for TIMIT are presented in Table 3 (2nd column). The triphone results are for models with 16 Gaussian mixtures (best performing triphones). The mixed-unit recogniser contained 151 syllable units, trained using a minimum of 60 tokens.

As can be seen in Table 3, longer-length acoustic units resulted in substantial gains in word accuracy. In fact, the relative reduction in WER achieved by going from triphones to a mixed-unit recogniser was 42%.

Table 3: Word accuracies (%), with a 95% confidence interval, on TIMIT and CGN.

Recogniser type	TIMIT	CGN
Triphone	91.9 ± 0.6	91.8 ± 0.6
Syllable-unit	93.5 ± 0.5	92.9 ± 0.6
Mixed-unit	95.3 ± 0.5	93.3 ± 0.6

4.2. CGN

The results for CGN are shown in Table 3 (3rd column). The triphone results are for models with 8 Gaussian mixtures (best performing triphones). The mixed-unit recogniser contained 94 syllable units, trained using at least 130 tokens.

In the case of the triphone recogniser, the results for CGN were of the same level as the results for TIMIT– regardless of the large difference in the test set perplexities. This might suggest that the acoustic perplexity of CGN is lower than in TIMIT. More importantly, however, using a mixed-unit recogniser yielded a more modest WER reduction (18% relative) with CGN.

Other studies have also failed to reach the kind of improvements gained on TIMIT. The absolute improvement in recognition accuracy obtained with mixed models in [3] was only 0.5%, although the comparison with [1] might not be fair due to [3] using a different type of phone-based recogniser. In [4], the gain obtained due to the inclusion of longer-length acoustic units depended heavily on the recognition task: for telephone numbers, the performance even decreased. This raises the question of how the different results can be explained, and what they can tell us about longer-length acoustic units when it comes to their capability of modelling long-term spectral and temporal dependencies in speech.

5. Analysing the differences

In this section, we attempt to explain why the performance gain obtained with a mixed-unit recogniser differs so much between recognition tasks. Since we only had access to TIMIT and CGN, we approached the problem by means of a detailed analysis of the experiments carried out on these two corpora. To that end, we investigated the differences between the further trained syllable units and the triphones that they were initialised with. However, before embarking on such an analysis, we first checked whether evident differences between the linguistic structure of TIMIT and CGN could explain the results.

5.1. Structure of the corpora

If the acoustic perplexity of CGN is lower than that of TIMIT, one would expect less gain from improved acoustic modelling. One way to obtain lower acoustic perplexity is through a higher proportion of polysyllabic words, which are intrinsically easier to recognise. However, it becomes obvious from Table 4 that the word structure of the two data sets is highly similar. Thus, there is no reason to believe that the baseline performance of CGN is difficult to improve upon due to a large proportion of long polysyllabic words.

We also checked for other differences between the corpora, such as the number of pronunciation variants and the durations of syllables. However, we were not able to identify linguistic or phonetic properties of the corpora that could possibly explain the differences in performance gain.

Table 4: Proportions of words with different numbers of syllables in TIMIT and CGN.

Number of syllables	TIMIT	CGN
1	63.1%	62.2%
2	22.7%	22.6%
3	9.3%	9.9%
4	3.5%	3.9%
≥ 5	1.4%	1.4%

5.2. Effect of further training

To investigate what happens when the longer-length units are trained further from the sequences of triphones used to initialise them, the degree to which the HMM states of the final, further trained syllable units differ from those of the initialised syllable units was examined. To this end, the distances between the probability density functions (pdf's) of the HMM states of the further trained syllable units and of the corresponding states of the initialised syllable units were calculated in terms of the Kullback-Leibler Distance (KLD) [8]. The KLD distributions for TIMIT and CGN are presented in Figure 1. The distributions differ from each other substantially, the KLDs generally being higher in the case of TIMIT. This implies that the further training affected the TIMIT units more than the CGN units. This is what one would expect, given the greater impact of the longer-length units on the recognition performance.

As the topologies of the concatenated triphones and the eventual syllable units are identical, there are two possibilities for explaining the larger impact of the further training on the TIMIT units. Either the boundaries of the syllable units with the largest KLD distances have shifted substantially, or the effect is due to the switch from the manually labelled phones to the further trained canonical representations of the syllable units. Since the syllable segmentations obtained through forced alignment did not show major differences, the issue of potential discrepancies between manual and canonical transcriptions was pursued further. We performed a new experiment on TIMIT, in which triphones were trained based on the canonical transcriptions of the uttered words. These 'canonical triphones' were then used to build a mixed-unit recogniser. The results of this experiment are shown in Table 5.

Table 5: Performance on TIMIT when using canonical transcriptions.

Recogniser type	Word acc. (%)
Triphone	96.0 ± 0.4
Mixed-unit	95.8 ± 0.4

Surprising as it may seem, the results obtained with the canonical triphones substantially outperform the results achieved with the manual labels; even the performance of the original mixed-unit system is significantly improved upon. The canonical triphones also outperform the new mixed-unit system, even though the difference in recognition performance is not statistically significant. The lack of improvement in recognition performance is reflected in smaller KLD distances between the initial and the further trained syllable units, as can be seen in Figure 2. Evidently, only a few syllables benefit from the further training, leaving the overall effect on the recognition performance negligible. These results are in line with results from other studies [3, 4], in which improvements due to longer-length acoustic units were smaller, and even deteriorations in recognition performance occurred.

The most probable explanation for the finding that the canonical triphones for TIMIT outperform the triphones trained using manual labels is the mismatch between the representations of speech during training and testing. While careful manual transcriptions yield more accurate acoustic models, the advantage of these models can only be reaped if the recognition lexicon contains a corresponding level of information about the pronunciation variation present in the speech [9]. Thus, it seems that at least part, if not all, of the performance gain obtained with further trained syllable units in the first set of experiments was due to the reduction of the mismatch between the representations of speech during training and testing.

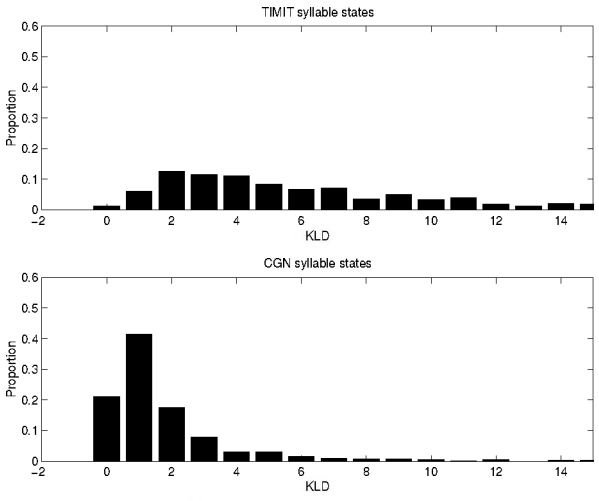

Figure 1: KLD distributions for the states of further trained syllable units in TIMIT and CGN.

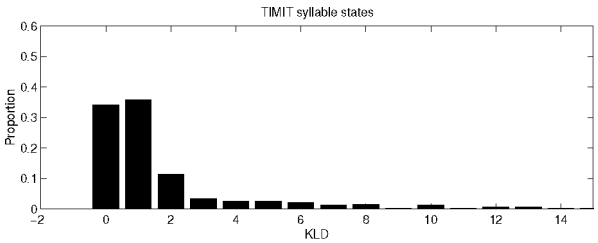

Figure 2: KLD distributions for the states of further trained syllable units in TIMIT when using canonical transcriptions.

At the time of writing this paper, we are performing new experiments on the CGN data, to investigate whether here, too, the improvement obtained with a mixed-unit recogniser was due to the removal of the mismatch between the representations of the training and test data. In general, the effect is expected to be smaller, if only because the CGN transcription protocol was based on a verification of the canonical transcription: transcribers were requested to change the canonical transcription only if a clearly different pronunciation variant had been realised by the speaker. As a consequence, the difference between the manual and canonical representations of the training data in CGN is much smaller than in TIMIT.

6. Discussion

The finding that the effect of longer-length acoustic units – which were expected to capture long-term coarticulation better than context-dependent phone models – might actually be negligible raises the question what else can be done to overcome the frequently cited limitations of phone models. We believe that retraining output pdf's does capture long-term coarticulation, but that this is not sufficient to capture the most important effects of pronunciation variation at the syllable level. Several authors – [10] in particular – have shown that, while syllables are seldom deleted completely, they do display considerable variation in the identity and number of the phonemes that best reflect their pronunciation. This type of variation can only be captured by designing more complex topologies, so that different variants can be modelled by different paths through the model.

Until now, changing the topology of triphone models to better represent pronunciation variation has met with limited success [11]. We believe that syllable units, which are often intrinsically longer than triphones, are subject to more variation. Therefore, we think that optimal topologies for syllable units could substantially improve the performance of LVCSR. Research is under way to investigate this issue for a larger set of read speech from CGN.

7. Conclusions

This paper contrasted recognition results obtained using longer-length acoustic units for Dutch read speech from a library for the blind with recognition results achieved on American English read speech from TIMIT. In both cases, substantial improvements over the performance of a triphone recogniser trained using manually labelled speech were obtained with a mixed-unit recogniser comprising syllable- and phoneme-length units. This may seem to corroborate the claim that properly initialised and further trained longer-length acoustic units capture a significant amount of pronunciation variation related to coarticulation spanning several phonemes. The KLD analysis carried out confirms that longer-length acoustic units may indeed capture some long-term coarticulation effects. However, detailed analysis of the results suggests that the effect of training syllable-sized units further is rather small if canonical representations of the syllables are initialised with triphone models trained on canonical transcriptions of the training corpus. We believe that the types of pronunciation variation that probably have the largest impact on recognition performance can only be captured by developing syllable units with a multi-path topology.

Acknowledgements

The research was carried out within the framework of the Interactive Multimodal Information eXtraction (IMIX) program, which is sponsored by Netherlands Organisation for Scientific Research (NWO).

References

[1] Sethy, A. and Narayanan, S., "Split-lexicon based hierarchical recognition of speech using syllable and word level acoustic units," in *Proc. ICASSP-2003,* Hong Kong, Apr 6-10, 2003, vol. 1, pp. 772-776.

[2] Hämäläinen, A., de Veth, J., and Boves, L., "Longer-Length Acoustic Units for Continuous Speech Recognition," in *Proc. EUSIPCO-2005*, Antalya, Turkey, Sep 4-8, 2005.

[3] Sethy, A., Ramabhadran B., and Narayanan, S., "Improvements in ASR for the MALACH project using syllable-centric models," in *Proc. IEEE ASRU-2003,* St. Thomas, US Virgin Islands, Nov 30 – Dec 4, 2003.

[4] Messina, R. and Jouvet D., "Context dependent "long units" for speech recognition," in *Proc. ICSLP-2004*, Jeju Island, Korea , Oct 4-8, 2004, pp. 645-648.

[5] TIMIT Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-505065, 1990.

[6] Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.P., Moortgat, M., and Baayen, H., "Experiences from the Spoken Dutch Corpus Project," in *Proc. LREC-2002*, Las Palmas de Gran Canaria, Spain, May 29-31, 2002, vol. 1, pp. 340-347.

[7] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., *The HTK Book (for HTK Version 3.2.1).* Cambridge University, Cambridge, UK, 2002.

[8] Kullback, S. and Leibler, R., "On information and sufficiency", *Annals of Mathematical Statistics*, 22:79-86, 1951.

[9] Wester, M., *Pronunciation Variation Modeling for Dutch Automatic Speech Recognition*. PhD thesis, University of Nijmegen, 2002.

[10] Greenberg, S., "Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation", *Speech Communication*, 29:159-176, 1999.

[11] Ostendorf, M. and Singer, H., "HMM topology design using Maximum Likelihood Successive State Splitting", *Computer Speech and Language,* 11:17-41, 1997.